POPS: A Software for Prediction of Population Genetic Structure Using Latent Regression Models Olivier François Eric Y. Durand

Flora Jay, Olivier François, Eric y Durand, Michael Gb Blum

Journal of Statistical Software

December 2015, Volume 68, Issue 9. doi: 10.18637/jss.v068.i09

POPS: A Software for Prediction of Population Genetic Structure Using Latent Regression Models

Flora Jay
Université Joseph Fourier

Olivier François
Université Joseph Fourier

Eric Y. Durand
Université Joseph Fourier

Michael G. B. Blum
Université Joseph Fourier

Abstract

The software **POPS** performs inference of population genetic structure using multilocus genotypic data. Based on a hierarchical Bayesian framework for latent regression models, **POPS** implements algorithms that improve estimation of individual admixture proportions and cluster membership probabilities by using geographic and environmental information. In addition, **POPS** defines ancestry distribution models allowing its users to forecast admixture proportion and cluster membership geographic variation under changing environmental conditions. We illustrate a typical use of **POPS** using data for an alpine plant species, for which **POPS** predicts changes in spatial population structure assuming a particular scenario of climate change.

Keywords: latent class regression models, mixture models, MCMC, population genetic structure, environmental covariates.

1. Introduction

Associations between population genetic structure and ecological variables have been frequently reported in the recent literature (Duminil *et al.* 2007; Aitken, Yeaman, Holliday, Wang, and Curtis-McLane 2008; Sork *et al.* 2010; Lee and Mitchell-Olds 2011). Predicting how changes in environmental conditions could impact this structure is important to several domains including molecular ecology, landscape genetics, conservation genetics or genetic epidemiology (e.g., Manel, Schwartz, Luikart, and Taberlet 2003; Storfer *et al.* 2006; Balding 2006; Balkenhol, Waits, and Dezzani 2009; Segelbacher *et al.* 2010).

Population genetic structure is commonly estimated by identifying *genetic clusters* defined as genetically divergent groups of individuals that arise from isolation of populations, and by computing individual membership probabilities for each genetic cluster (Davies, Villablanca, and Roderick 1999; Pritchard, Stephens, and Donnelly 2000). Because genetic ancestry can be shared among several clusters, another way to estimate population structure is to infer *admixture proportions* representing the relative contributions of distinct ancestral populations to a genome. Identifying genetic clusters based on multilocus genetic data can be viewed as an instance of unsupervised learning based on multivariate categorical variables. The first efforts to infer genetic clusters and individual admixture proportions using Bayesian modeling date back to the introduction of the computer programs **STRUCTURE** and **PARTITION** (Pritchard *et al.* 2000; Dawson and Belkhir 2001). Although **STRUCTURE** algorithms and their derivatives were widely used to study the influence of landscape features on evolutionary processes, these approaches did not incorporate information on environmental variables.

In this article, we introduce **POPS**, a software that implements Bayesian clustering algorithms based on genetic, geographic and environmental variables. The principle of **POPS** is that individuals sharing similar environmental conditions and geographically close to each other are also likely to share genetic ancestry. To achieve this objective **POPS** assigns individuals or genes to genetic groups after modeling the effects of geography and environment on individual membership and admixture proportions. **POPS** is based on latent regression models that consider individual ancestry as a hidden response variable regressed on geographical and environmental covariates (e.g., Bandeen-Roche, Miglioretti, Zeger, and Rathouz 1997; Chung, Flaherty, and Schafer 2006; Linzer and Lewis 2011). Hidden regression models are used to obtain predictions of cluster membership and admixture proportions from geographic and environmental variables. In the context of climate change, **POPS** can be used to forecast changes in population genetic structure of species in response to global warming (Jay *et al.* 2012).

POPS is implemented in the C++ programming language and can be run from a command-line engine or a graphical user interface. The program takes input data files in a format compatible with existing clustering algorithms like **STRUCTURE** (Pritchard *et al.* 2000) and **TESS** (Chen, Durand, Forbes, and François 2007). **POPS** returns textual and graphical results of inferred and predicted membership probabilities and admixture coefficients, allele frequencies in the estimated clusters, and values of criteria for model selection. In Sections 2 and 3, we describe the hierarchical models used by **POPS** and their implementation using Markov chain Monte Carlo (MCMC) algorithms. In Section 4, we explain how to use the software from its graphical user interface and from its command-line engine. In Section 5, we illustrate the use of **POPS** on an example data set and show the main features of our program.

2. Models

We consider a data matrix, x, with N rows and L columns that records genotypic data for a sample of N individuals genotyped at L loci. For haploid species, each entry of x corresponds to the occurrence of a given allele encoded as a categorical variable. In polyploid species there are $A \times N$ ($A \geq 2$) rows instead of N rows in the data matrix. Each copy corresponds to one of the A chromosomes carried by an organism. To each sampled individual corresponds a geographic sample site where coordinates are recorded. In addition to the

genetic and geographical data, we assume that a set of environmental covariates are measured at each sample site. We denote by $\tilde{X}^{\mathrm{S}}$ the geographic coordinates of a site, and by $\tilde{X}^{\mathrm{E}}$ the subset of measured environmental variables. We denote by $\tilde{X}$ the $N \times (D+1)$ design matrix containing the value 1 in its first column plus D geographical and environmental covariates in its subsequent columns.

2.1. Objectives of POPS models

The models implemented in the program **POPS** fall into two main classes according to whether or not genetic admixture is considered. A model without admixture assumes that each individual originates from a unique ancestral group whereas a model with admixture assumes that individuals may share ancestry from more than one source population. In statistical terms, individual membership to a genetic cluster and admixture proportions correspond to non-observed random variables. For readers not familiar with Bayesian inference of population structure, we recommend to read Pritchard *et al.* (2000) where the software **STRUCTURE** is introduced and the standard Bayesian model of population genetic structure is carefully explained.

Assuming that there are K ancestral groups – also called genetic clusters – models without admixture infer latent group labels corresponding to the membership of each individual, z_i, of one of the K clusters. As with the program **STRUCTURE**, **POPS** estimates a matrix of ancestral allele frequencies $p = (p_{k\ell j})$, $k = 1, \ldots, K$, $\ell = 1, \ldots, L$ and $j = 1, \ldots, J_\ell$. Each entry of p corresponds to the frequency of allele j at locus ℓ in population k, and J_ℓ is the number of distinct alleles observed at locus ℓ.

Models with admixture suppose that each individual genome shares ancestry from K ancestral clusters. Admixture models estimate a matrix, q, where each element, q_{ik}, $i = 1, \ldots, N$, $k = 1, \ldots, K$, corresponds to the proportion of individual i's genome that originates from cluster k. In admixture models, **POPS** estimates cluster labels, $z_i^{(\ell,a)}$, for each allele copy, $a = 1, \ldots, A$, at each locus, $\ell = 1, \ldots, L$.

POPS provides estimates and predictions for the hidden variables z and q. Compared to existing software packages, the basic principle of **POPS** is that the inclusion of geographical and environmental variables improves the accuracy of inference when there is genuine association between those variables and population genetic structure. Using latent correlation models and posterior predictive analysis, **POPS** additionally enables building geographic maps of ancestry distribution under various scenarios of environmental change.

2.2. Models without admixture

Let us denote by $x_i = \left(x_i^{(\ell,a)}\right)$ the multilocus genotype obtained by concatenating genotypes at all loci for individual i. Models without admixture are based on a refinement of latent class models that incorporates geographical and environmental covariates in mixture models. In latent class models, the data are modeled using mixtures of K class-specific distributions

$$\mathsf{P}(x_i) = \sum_{k=1}^{K} \mathsf{P}(z_i = k)\mathsf{P}(x_i | z_i = k)\,, \quad i = 1, \ldots, N. \tag{1}$$

In **POPS**, $\mathsf{P}(x_i | z_i = k)$ integrates over the allele frequencies p, and the conditional distribution

is defined as in the **STRUCTURE** model (Pritchard *et al.* 2000)

$$\mathsf{P}(x_i|z,p) = \prod_{\ell=1}^{L}\prod_{a=1}^{A} p_{z_i \ell x_i^{(\ell,a)}}. \tag{2}$$

In other words, the allele frequencies at each locus are assumed to be independent within each ancestral group, and the distribution of allele counts at each locus is assumed to be multinomial with frequency $p_{k\ell\cdot}$.

POPS incorporates knowledge on geographical and environmental covariates, $\tilde{X}$, by considering the following distribution

$$\mathsf{P}(x_i|p,\tilde{X}) = \sum_{k=1}^{K} \mathsf{P}(z_i = k|\tilde{X})\mathsf{P}(x_i|z_i = k, p, \tilde{X}).$$

More specifically **POPS** considers a particular class of latent class models called concomitant-variable latent class or latent class feed-forward models (Dayton and Macready 1988; Vermunt and Magidson 2003). Such models were implemented in the R (R Core Team 2015) package **poLCA** (Linzer and Lewis 2011). They assume that the covariates have no influence on the distribution of the data in a given class, i.e., the right term in the previous sum can be simplified as

$$\mathsf{P}(x_i|z_i = k, p, \tilde{X}) = \mathsf{P}(x_i|z_i = k, p).$$

In **POPS**, the conditional probabilities $\mathsf{P}(z_i = k|\tilde{X})$ are computed using a probit regression model (Jay, François, and Blum 2011). To emphasize geographical and environmental covariates, we denote by $\tilde{X}^{\mathrm{S}}$ the subset of geographic covariates and by $\tilde{X}^{\mathrm{E}}$ the subset of environmental covariates. In the probit model, the cluster K is considered as a *reference* cluster, and all regression coefficients are estimated with respect to this reference. The probit model considers vectors of unobserved continuous variables $c_i = (c_{i,1}, \ldots, c_{i,K-1})$ as response variables in $K-1$ regression equations (Albert and Chib 1993)

$$c_{i,k} = \tilde{X}_i^{\mathrm{E}}\beta_k^{\mathrm{E}} + f(\tilde{X}_i^{\mathrm{S}})\beta_k^{\mathrm{S}} + \epsilon_{i,k}, \tag{3}$$

where

$$\begin{bmatrix} \epsilon_{i,1} \\ \vdots \\ \epsilon_{i,K-1} \end{bmatrix} \sim \mathcal{N}(0, \mathrm{Id}_{K-1}),$$

Id_{K-1} is the identity matrix, β_k^{E} and β_k^{S} are vectors of regression parameters, and f is a polynomial function of degree lower than 3. The function f represents a trend surface for the unobserved variables $c_{i,k}$. Its introduction is motivated by the definition of admixture models for which the estimation of individual ancestry coefficients is performed by using trend surfaces and universal kriging models (see next section and Durand, Jay, Gaggiotti, and François 2009). The variable z_i can be obtained from the vector c_i as follows

$$z_i = \begin{cases} K & \text{if } \max_{k'} c_{i,k'} < 0 \\ k & \text{if } \max_{k'} c_{i,k'} \geq 0 \text{ and } \max_{k'} c_{i,k'} = c_{i,k}\,. \end{cases} \tag{4}$$

Since the probability of having more than one cluster label maximizing $c_{i,k}$ is equal to zero, the definition of z_i is unambiguous. Note that equivalent clustering solutions can be obtained from the models after any permutation of group labels.

Finally, the posterior distribution in the model without admixture can be written as

$$\mathsf{P}(z, p, \beta | x, \tilde{X}) \propto \mathsf{P}(x|z, p)\mathsf{P}(z|\tilde{X}, \beta)\mathsf{P}(\beta)\mathsf{P}(p), \tag{5}$$

where $\mathsf{P}(p)$ and $\mathsf{P}(\beta)$ are prior distributions. The prior distribution on allele frequencies within each ancestral group is a Dirichlet distribution

$$p_{k\ell \cdot} \sim \mathcal{D}(\lambda, \ldots, \lambda), \tag{6}$$

where λ is set to 1 by default. Note that other values of λ can be chosen by a **POPS** user.

2.3. Models with admixture

Models with admixture assume that the genome of each individual originates from K ancestral populations. Since there is a latent class variable for each allele at each locus, admixture models can be considered as extensions of latent class models. In the admixture models implemented in **STRUCTURE**, Equation 1 extends to

$$\mathsf{P}(x_i|p, q_{i\cdot}) = \prod_{\ell=1}^{L} \prod_{a=1}^{A} \sum_{k=1}^{K} \mathsf{P}(z_i^{(\ell,a)} = k|q_{ik})\ \mathsf{P}(x_i^{(\ell,a)}|z_i^{(\ell,a)} = k, p), \tag{7}$$

where q_{ik} denotes the admixture coefficient of individual i in population k.

POPS incorporates geographical and environmental covariates in admixture models as in latent class feed-forward models, which implies

$$\mathsf{P}(p, q|x, \tilde{X}) \propto \mathsf{P}(x|p, q)\mathsf{P}(q|\tilde{X})\mathsf{P}(p), \tag{8}$$

where $\mathsf{P}(x_i|p, q)$ is given by Equation 7 and $q_{i\cdot}$ depends on the covariates $\tilde{X}$ via the hyperparameters $\alpha_{i\cdot}$ as described in the next paragraph.

The admixture coefficients are sampled from a Dirichlet distribution

$$q_{i\cdot} \sim \mathcal{D}(\alpha_{i1}, \cdots, \alpha_{iK}), \tag{9}$$

where the parameters, α_{ik}, are proportional to the expected amount of individual admixture from each ancestral group. The geographical and environmental covariates $\tilde{X}$ are incorporated by considering the parameter α as a response variable in a multivariate latent regression model (Durand *et al.* 2009). **POPS** implements the following log-normal model

$$\log(\alpha_{ik}) = \tilde{X}_i^{\mathrm{E}}\beta_k^{\mathrm{E}} + f(\tilde{X}_i^{\mathrm{S}})\beta_k^{\mathrm{S}} + y_{ik}, \tag{10}$$

where y_{ik} is a zero-mean conditional autoregressive Gaussian model (CAR; Besag 1975; Ripley 1981; Vounatsou, Smith, and Gelfand 2000). In Equation 10 the second term represents a spatial trend that accounts for broad-scale spatial patterns. The third term corresponds to a spatially autocorrelated residual that accounts for local effects. The first term measures the effect of environmental covariates once spatial effects are removed (Lichstein, Simons, Shriner, and Franzreb 2002). In the CAR model, the distribution of y_{ik} depends on the values of y_{jk} at neighboring sites. The model is conditionally Gaussian with mean

$$\mathsf{E}[y_{ik}|y_{jk}, j \neq i] = \rho_k \sum_{j \neq i} w_{ij} y_{jk}, \tag{11}$$

and variance

$$\mathsf{VAR}(y_{ik}|y_{jk}, j \neq i) = \sigma_k^2, \tag{12}$$

where ρ_k is the magnitude of the spatial neighborhood effect in cluster k, w_{ij} are weights that determine the neighborhood of i and the relative effect of j on i. The parameter σ_k^2 is a variance parameter for the cluster k. The neighborhood is obtained from a Dirichlet tessellation built on sample site coordinates (François, Ancelet, and Guillot 2006). The weights are functions of the great-circle distance between sites, d_{ij},

$$w_{ij} = \exp\left(-\frac{d_{ij}}{\theta}\right), \tag{13}$$

where θ is equal to the mean value of great-circle distance between sample sites.

The posterior distribution of the multidimensional parameter (z, p, β, q, α) can be obtained as

$$\mathsf{P}(z, p, \beta, q, \alpha|x, \tilde{X}) \propto \mathsf{P}(x|z, p)\mathsf{P}(z|q)\mathsf{P}(q|\alpha)\mathsf{P}(\alpha|\tilde{X}, \beta)\mathsf{P}(\beta)\mathsf{P}(p). \tag{14}$$

The prior distribution on p is given by Equation 6, and the conditional distribution $\mathsf{P}(z|q)$ is given by

$$\mathsf{P}(z_i^{(\ell,a)} = k|q) = q_{ik},\ i = 1, \ldots, N,\ a = 1, \ldots, A,\ \ell = 1, \ldots, L.$$

2.4. Label switching

In this section, we briefly discuss label switching that is a common issue for **STRUCTURE**-like models. Label switching refers to the invariance of the likelihood function under relabeling of the mixture components. The likelihood is unchanged when permuting the labels (Stephens 2000). Because of this invariance to permutation, the marginal posterior distribution of an individual's membership coefficient, z_i, is given by $\mathsf{P}(z_i = k|x, \tilde{X}) = 1/K$, for $k = 1, \ldots, K$. Label switching in **POPS** can be addressed using the software **CLUMPP** (Jakobsson and Rosenberg 2007) that provides alignments of distinct clustering results, and allows users to compare these results. Once distinct clustering solutions are aligned with **CLUMPP**, we recommend to use a model averaging approach based on the outputs of several MCMC runs. Averaging the prediction of membership or admixture coefficients over several posterior regions obtained with different MCMC outputs improve prediction performance.

3. Inference and posterior predictive simulations

In this section we describe the MCMC algorithms implemented in **POPS** to perform parameter inference, sample from "modes" of the posterior distribution, deal with label switching issues (identifiability), and use posterior predictive simulations to make predictions using new values of the geographical and environmental covariates.

3.1. Models without admixture

To sample from the posterior distribution $\mathsf{P}(z, p, \beta|x, \tilde{X})$, **POPS** implements MCMC algorithms using Gibbs sampling updates.

Updating p. To update the allele frequencies p, we consider the same Gibbs sampler updating step as in the software **STRUCTURE** (Pritchard *et al.* 2000). It is performed by simulating allele frequencies as follows

$$p_{k\ell\cdot}|x, z \sim \mathcal{D}(\lambda + n_{k\ell 1}, \ldots, \lambda + n_{k\ell J_\ell}), \tag{15}$$

where $n_{k\ell j}$ denotes the number of copies of allele j in population k at locus ℓ, $k = 1, \ldots, K$, $\ell = 1, \ldots, L$, $j = 1, \ldots, J_\ell$.

Updating z. Since the cluster label z can be obtained from the latent variable c in a deterministic fashion, z and c are updated simultaneously. Using Bayes' formula, the conditional distribution of (c, z) can be written as

$$\mathsf{P}(c, z|\beta, p, x, \tilde{X}) \propto \mathsf{P}(x|p, z)\mathsf{P}(c|\beta, \tilde{X})\mathsf{P}(z|c).$$

Samples from the above conditional distribution are generated using the following rejection algorithm:

Step 1. For each i, generate c_i from the regression Equation 3 and determine z_i with the rule described in Equation 4.

Step 2. Given $z_i = k$ from Step 1, accept the pair (c_i, z_i) with probability

$$\frac{\mathsf{P}(x_i|p, z_i = k)}{\max_k \mathsf{P}(x_i|p, z_i = k)},$$

otherwise return to Step 1. The probability $\mathsf{P}(x_i|p, z_i = k)$ can be computed from Equation 2.

Updating β. **POPS** uses a diffuse prior distribution $\beta \sim \mathcal{N}(0, B^{-1})$ with $B = 0$. The Gibbs sampler proceeds by updating values of β using its conditional distribution (Albert and Chib 1993)

$$\beta_k|c \sim \mathcal{N}(V\Omega^\top c_{\cdot k},\ V), \tag{16}$$

where $V = (\Omega^\top\Omega)^{-1}$, and Ω is the concatenation of the geographical $f(\tilde{X}^{\mathrm{S}})$ and environmental $\tilde{X}^{\mathrm{E}}$ variables.

3.2. Models with admixture

To perform inference in admixture models, **POPS** uses a hybrid MCMC algorithm. Unless specified otherwise, updates are done using Gibbs samplers. Updates of p, q and z are similar to those used in the algorithm of **STRUCTURE** (Pritchard *et al.* 2000). The updates of $(\alpha, \beta, \sigma^2, \rho)$ are similar to those used in the algorithm of **TESS** (Durand *et al.* 2009).

Updating p. Given x and z, p is simulated according to Equation 15 that is given for the model without admixture.

Updating q. For an individual i, the admixture coefficients are sampled from a Dirichlet distribution

$$q_{i\cdot}|x, z, \alpha \sim \mathcal{D}(\alpha_{i1} + m_{i1}, \ldots, \alpha_{iK} + m_{iK}) \tag{17}$$

where m_{ik} is the number of allele copies that originated from population k.

UPDATING z. A cluster label $z_\ell^{(i,a)}$ is simulated independently for each triplet (i, a, ℓ) from

$$\mathsf{P}(z_i^{(\ell,a)}|p,q,x) = \frac{q_{ik}\mathsf{P}(x_i^{(\ell,a)}|p,\ z_i^{(\ell,a)}=k)}{\sum_{k'=1}^{K} q_{ik'}\mathsf{P}(x_i^{(\ell,a)}|p, z_i^{(\ell,a)}=k')}. \tag{18}$$

UPDATING (α, y). The parameters (α_{ik}, y_{ik}) are updated for each pair (i,k) using a Metropolis-Hastings algorithm. For each $(i_0,k_0) \in \{1,\ldots,N\}\times\{1,\ldots,K\}$, a new value $y^*_{i_0k_0}$ is sampled from the Gaussian conditional probability distribution

$$y^*_{i_0k_0}|\rho,\sigma^2,y_{ik},i\neq i_0 \ \sim\ \mathcal{N}(\rho\sum_{j\neq i_0} w_{i_0j}\ y_{jk_0},\sigma^2). \tag{19}$$

The latent variable $y^*_{i_0k_0}$ is accepted with probability

$$\min\left(1,\ \frac{\Gamma\left(\sum_{k=1}^{K}\alpha^*_{i_0k}\right)\Gamma\left(\alpha_{i_0k_0}\right)}{\Gamma\left(\sum_{k=1}^{K}\alpha_{i_0k}\right)\Gamma\left(\alpha^*_{i_0k_0}\right)} q_{i_0k_0}^{\alpha^*_{i_0k_0}-\alpha_{i_0k_0}}\right) \tag{20}$$

where $\alpha^*_{i_0k_0} = \exp\left(\tilde{X}^E_{i_0}\beta^E_{k_0} + f(\tilde{X}^S_{i_0})\beta^S_{k_0} + y^*_{i_0k_0}\right)$, $\alpha_{i_0k_0}$ is the current value of the parameter, and Γ denotes the Gamma function. If $y_{i_0k_0}$ is updated to $y^*_{i_0k_0}$, then $\alpha_{i_0k_0}$ is updated to $\alpha^*_{i_0k_0}$.

UPDATING β. **POPS** uses a diffuse prior distribution on β, $\beta\sim\mathcal{N}(0,B^{-1})$, with $B=0$, so that the conditional posterior distribution on β is given by

$$\beta_k|\alpha,\rho_k,\sigma_k,\tilde{X}\ \sim\ \mathcal{N}(V\Omega^\top(\mathrm{Id}-\rho_k W)\log(\alpha_k),\sigma_k^2 V) \tag{21}$$

where $V=\left(\Omega^\top(\mathrm{Id}-\rho W)\Omega\right)^{-1}$, Id is the identity matrix, $\Omega=(\tilde{X}^{\mathrm{E}}, f(\tilde{X}^{\mathrm{S}}))$, and $W=(w_{ij})$.

UPDATING σ^2. **POPS** uses a Gamma distribution for updating the hyperprior parameter $\phi_k = 1/\sigma_k^2$ for each k in $\{1,\ldots,K\}$ given by

$$\phi_k|y\sim\mathrm{Ga}\left(\frac{N}{2},\frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq i} w_{ij}\ y_{ik}\ y_{jk}\right), \tag{22}$$

where $\mathrm{Ga}(a,b)$ denotes the Gamma distribution with shape a and rate b.

UPDATING ρ. Let $e=(e_1,\ldots,e_N)$ be the N eigenvalues of the weight matrix W, and $e_{\max}$ the largest eigenvalue. **POPS** uses a uniform hyperprior distribution in the range $(0, e_{\max}^{-1})$ on ρ. **POPS** updates ρ_k independently for each k by using a Metropolis-Hastings step. A new value ρ^* of ρ_k is proposed from a Gaussian random walk with a fixed variance equal to 0.05. **POPS** rejects the proposed value when the value is not within the interval $(0, e_{\max}^{-1})$. Otherwise, the program accepts it with probability

$$\min\left(1,\ \prod_{i=1}^{N}\left(\frac{1-\rho^* e_i}{1-\rho_k e_i}\right)^{1/2}\ \exp\left(-\frac{1}{2\sigma_k^2}\ y_{ik}\sum_{j=1}^{N} w_{ij}\ y_{jk}\ (\rho_k-\rho^*)\right)\right). \tag{23}$$

3.3. Posterior predictive simulations and model selection

An important feature of **POPS** is that the inferred model can be used for predicting cluster membership and admixture proportions given new values of environmental and spatial covariates. In models without admixture, membership coefficients can be obtained by sampling latent cluster labels using Equations 3 and 4, where regression coefficients are generated from their posterior distribution. For each individual, the relative frequency of predicted cluster labels provides a predicted membership probability for each cluster. Similarly in admixture models, admixture coefficients can be predicted using new values of the environmental and geographical covariates. Predicted admixture coefficients can be obtained by sampling regression coefficients from the posterior distribution, and by simulating admixture coefficients using Equations 9 and 10.

3.4. Addressing label switching and multimodality

To address label switching and multimodality, we use the following post-processing methods for the MCMC runs. First, we run several MCMC runs and we select a subset of runs minimizing the DIC (Spiegelhalter, Best, Carlin, and Linde 2002). Because of multimodality, MCMC runs may visit regions of the posterior distribution that correspond to distinct modes of the posterior distribution and are truly distinct solutions (Jakobsson and Rosenberg 2007). By selecting runs with the lowest DIC values, we discard runs that visit less interesting regions of the posterior distribution.

In **STRUCTURE** MCMC runs, it is generally observed that labels do not switch during a single MCMC run. When we run the MCMC algorithm several times, the replicates that we select using the DIC generally provide comparable membership or admixture coefficients after permutation of cluster labels (Jakobsson and Rosenberg 2007). We use the program **CLUMPP**, which implements an algorithm that deals with label switching by looking for an optimal alignment of estimates obtained from distinct MCMC runs. Then, we average the results over different MCMC runs in order to improve the prediction of cluster memberships and admixture coefficients.

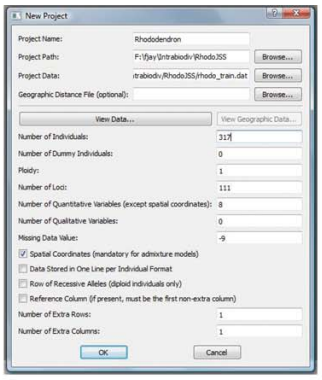
4. Using POPS

POPS is a free software implemented in the C++ programming language, and the graphical user interface uses the library Qt (Qt Project 2015). In this section, we briefly describe its graphical user interface, and we explain the main instructions of the command-line engine.

4.1. POPS graphical user interface (GUI)

To start an analysis using the **POPS** GUI, the user must create a project by loading a file containing genetic data, and spatial and environmental data (Figure 1). The data format required by **POPS** is similar to **STRUCTURE** or **TESS** formats. By default the genotypes should be stored on one row for an haploid individual, and two rows for a diploid individual. Other formats compatible with **STRUCTURE** and **TESS** can also be specified. In addition to geographic or environmental data, there can be additional rows or columns present in the input file for informational or other purposes. Columns must be stored in a predefined order: (1) extra columns (optional, e.g., identifier for samples, population labels), (2) qualita-

A



B

	1	2	3	4	5	6	7	8	9	10	11	12	
1	Ind	prcpavgy	tmaxavgty	tminavgty	asp200	slp200	topo	twi200s	dem200ibd	longitude	latitude	AAT_CAC_63.6	AAT
2	E21-1	255.11	7.21	-0.28	298.17	14.11	385	-0.89	1552	11.29	47.61	0	
3	E21-2	255.11	7.21	-0.28	298.17	14.11	385	-0.89	1552	11.29	47.61	0	
4	E21-4	255.11	7.21	-0.28	298.17	14.11	385	-0.89	1552	11.29	47.61	0	
5	E23-1	216.01	6.19	-1.19	158.85	15.84	327	-1.07	1777	11.93	47.64	0	
6	E23-2	216.01	6.19	-1.19	158.85	15.84	327	-1.07	1777	11.93	47.64	0	
7	E23-4	216.01	6.19	-1.19	158.85	15.84	327	-1.07	1777	11.93	47.64	1	
8	E25-1	309.16	6.37	-0.57	102.46	10.38	261	-0.73	1739	12.33	47.71	0	
9	E25-2	309.16	6.37	-0.57	102.46	10.38	261	-0.73	1739	12.33	47.71	0	

Figure 1: A: Creation of a new project. The user has to input a data file and data information. B: View of the data file in **POPS**. Environmental data are highlighted in green, coordinates in blue, and genetic markers in white.

tive environmental covariates (optional), (3) quantitative environmental covariates (optional), (4) longitude and latitude (required in admixture models), (5) genetic markers (required).

Users must specify the following parameters:

- model with or without admixture,
- degree of the spatial trend surface (0, 1, 2, or 3),
- range of values for the maximal number of clusters K,
- number of runs to launch for each value of K,
- number of sweeps of each run (total number of steps and number of burn-in iterations).

Optional parameters can be specified, including hyperparameters for admixture models: the scale parameter θ and initial CAR variances (one value for all clusters).

4.2. Outputs

The program **POPS** provides textual and graphical summaries of MCMC runs, including estimates of model parameters, log-likelihood histories of each run for convergence diagnostics, DIC values for model selection, and visualizations of vectors of posterior probabilities for cluster labels (models without admixture), or matrices of admixture coefficients (admixture models). These graphical outputs use barplot representation of ancestry coefficient matrices, as commonly displayed by other ancestry estimation programs. When geographic information

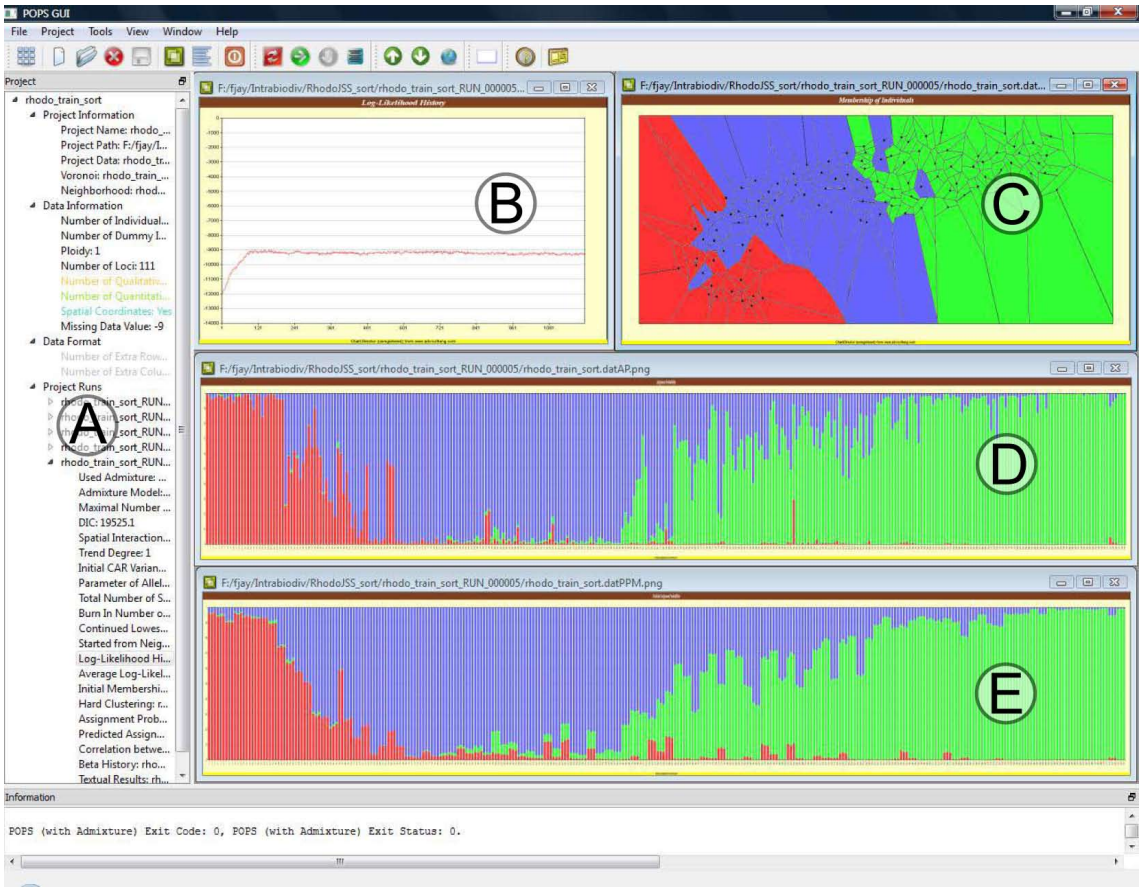


Figure 2: Analysis of a *Rhododendron ferrugineum L.* data set using the **POPS** GUI. Graphical results are displayed for one run. A: a tree widget from which the user has access to project and data information, run options, textual and graphical outputs. All output results can be displayed by double-clicking on the appropriate items in the tree widget B: Log-likelihood history of the run. C, D, E: 3 graphical outputs, where colors correspond to clusters. C: Each plant is assigned to one cluster (for which the admixture coefficient is maximal) and assignments are displayed where each black point represents a sample site. D: Admixture coefficients estimated for each sample. E: Admixture coefficients predicted from environmental data for each sample. In D and E, individuals are represented as vertical lines partitioned into segments corresponding to the fraction of their genomes assigned/predicted to each genetic cluster.

is provided, **POPS** displays *hard-clustering* assignments where each individual is assigned to the cluster for which the membership or admixture coefficient is maximal. An additional feature of **POPS** is to compute predictions of population genetic structure for new environmental data. Supplementary scripts are provided on the **POPS** web page to display actual or predicted population genetic structures on geographic maps using the **fields** package for R (Nychka, Furrer, and Sain 2015). Finally, correlation coefficients evaluating the correlation between matrices of inferred and predicted coefficients are computed and returned to users for facilitating the interpretation of results.

Output results are accessible by double-clicking on the corresponding items in the tree widget

(Figure 2A). In addition, the "Summarize Runs" button allows the user to quickly access the summaries of all runs (K, trend degree, model used, DIC, average log-likelihood, correlation score). **POPS** also provides a tool to export runs to **CLUMPP** that can average membership and admixture coefficients estimated from multiple MCMC runs (Jakobsson and Rosenberg 2007). Finally, the summary window can be used to perform predictions. To predict membership or admixture coefficients for new values of covariates, the user needs to load a file containing new data and choose runs that will be used for predictions. Textual and graphical results of predicted coefficients are computed for each selected run, and are accessible from the tree widget (Figure 2A). Prediction outputs from independent runs can also be exported to **CLUMPP**.

4.3. POPS command-line options

POPS is based on two command-line engines: one for models without admixture, and the other for models with admixture. We describe the main commands for both programs. When there are no options given to **POPS**, it will show its typical usage and exit. In the following the main **POPS** options are given. They can be specified in any order.

Required parameters:

`-F` File name of input data file.
`-N` Number of individuals.
`-A` Ploidy (`1` = haploid, `2` = diploid, ...).
`-L` Number of loci.
`-K` Maximal number of clusters.
`-T` Degree of trend: `-1`: No spatial coordinates in data file.
`0`: Spatial coordinates present but not used.
`1`, `2`, `3`: Degree 1, 2, or 3.
`-D` Parameter of Dirichlet allele frequency model.
`-S` Total number of sweeps of MCMC.
`-B` Burn-in number of sweeps of MCMC.
`-P` Spatial interaction parameter (for admixture models only).

Optional parameters:

`-XL` Number of qualitative variables.
`-X` Number of quantitative covariates (other than spatial coordinates).
`-r` Number of extra rows in data file.
`-c` Number of extra columns in data file.
`-i` Folder name of input data file (default: current folder).
`-o` Folder Name of output result files (default: current folder).
`-orun` Suffix to append to output result file names, e.g., a run number (`-orun1` or `-orun0001`) or a specific run name (`-orunAdm002`).
`-sp` Special data format: 1 individual = 1 row (`-spy`: yes, `-spn`: no, default). Execute `pops|more` and `popsAdm|more` or see **POPS** manual for extra options.

The command `pops` (or `pops.exe` on the Windows operating system) runs models without admixture, whereas `popsAdm` (or `popsAdm.exe`) runs admixture models.

For example, suppose that the data set is stored in a file named `example.txt` in a folder `Example`. This file is provided with the software. The data contain 268 haploid individuals (`-N268 -A1`) genotyped at 86 loci (`-L86`), without any environmental data (`-XL0 -X0`), no

spatial coordinates (`-T-1`), 1 extra row `-r1` and 4 extra columns `-c4`. Assuming there are at most 3 clusters (`-K3`), we set the parameter of the Dirichlet allele frequency model to 1.0 (`-D1.0`). To run the MCMC algorithm for a total of 1,000 sweeps (`-S1000`) with the first 200 sweeps discarded as burn-in period (`-B200`), we use the following command

```
$ pops -Fexample.txt -N268 -A1 -XL0 -X0 -T-1 -L86 -K3 -D1.0 -S1000 -B200 \
> -r1 -c4 -iExample -oExample -orun001
```

Output results are stored in a directory `Example` (`-oExample`) and the string `_RUN001_` will be appended to the output names (`-orun001`).

Assume now that the file `example.txt` contains an extra column (identifier for samples, `-c1`), 1 quantitative covariate (temperature `-X1`), and 2 columns for longitude and latitude. To run **POPS** with the same run parameters as before but using the covariates, and setting the degree of the trend surface to 1 (`-T1`), the command is

```
$ pops -Fexample.txt -N268 -A1 -XL0 -X1 -T1 -L86 -K3 -D1.0 -S1000 -B200 \
> -r1 -c1 -iExample -oExample -orun002
```

To run the **POPS** admixture model using the same data and the same parameters as shown in the preceding command, we additionally specify the default value of the spatial interaction parameter (`-P0.6`)

```
$ popsAdm -Fexample.txt -N268 -A1 -XL0 -X1 -T1 -L86 -K3 -D1.0 -S1000 \
> -B200 -P0.6 -r1 -c1 -iExample -oExample -orun003
```

5. Examples

In this section, we illustrate several features of **POPS** using a data set of 377 individuals from plant species *Rhododendron ferrugineum L.*, with each individual genotyped at 111 loci (INTRABIODIV database, Gugerli *et al.* 2008). *Rhododendron ferrugineum L.* is a small and evergreen shrub present in European mountains. With the genetic data, we consider geographic and environmental covariates consisting of latitude, longitude, average minimum and maximum annual temperatures, average precipitations, and an additional set of topographic variables measured at each sampled site.

5.1. Estimating population genetic structure

A subset of 60 individuals were randomly chosen among the 377 samples to constitute a test set. We ran **POPS** on the remaining 317 plants, and we reported the results from a single run using the admixture model with $K = 3$ clusters. The run used 1,000 sweeps following a burn-in period of 200 sweeps. The log-likelihood function increased quickly during the run, and then reached a stationary state (Figure 2B). Admixture coefficients were estimated for each individual and displayed in the graphical user interface in Figure 2D. Hard-clustering assignments are displayed on the tessellation built from the sample sites locations (Figure 2C). The map shows that the 3 inferred genetic clusters correspond to three well-separated geographical regions, in the southwestern region (red cluster), central region (blue cluster) and

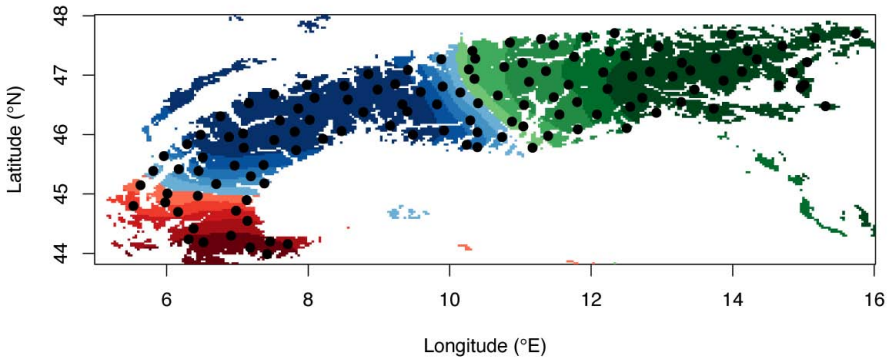


Figure 3: Admixture coefficients estimated for *Rhododendron ferrugineum L.* are displayed spatially on a map of the European Alps. The map is computed using an R script based on kriging methods and provided with **POPS**. Only coefficients greater than 0.5 are displayed.

northeastern region (green cluster). R scripts based on kriging methods allow us to display admixture coefficients spatially (Figure 3). Though substantial admixture occurs within contact zones between clusters, the results are consistent with hard-clustering assignments. To evaluate the improvement on the estimation of population genetic structure obtained with models using environmental covariates, we computed DIC for models that do not use any environmental information. We find that DIC decreases from 19565 to 19525 when adding environmental covariates.

5.2. Predicting population genetic structure based on covariates

POPS can test if a set of covariates included in a model is useful to predict population genetic structure by computing the correlation between the estimated admixture coefficients and admixture coefficients predicted from the geographical and environmental covariates (Figures 2D and E). A correlation score of ≈ 0.96, reported in the tree widget, indicates that prediction from environmental variables is accurate. When we use **POPS** to predict admixture coefficients from the covariates of 60 individuals contained in the test data set (and not used for inference), the correlation score is equal to 0.85 (Figure 4).

5.3. Forecasting population genetic structure under changes

According to the Intergovernmental Panel on Climate Change, environmental conditions may change drastically during the coming century (Intergovernmental Panel on Climate Change 2007). Temperatures are predicted to rise by 1.8 to 4 °C, depending on the IPCC expert projection. Precipitations are also likely to increase in several regions. These changes are now acknowledged to have an impact on species distributions and there has been increasing evidence of species' range shifts due to climate change (Parmesan and Yohe 2003). **POPS** provides a framework to investigate and forecast modifications in population genetic structure in response to environmental changes. We used **POPS** to forecast changes in population genetic structure of the species *Rhododendron ferrugineum L.* under a 2 °C temperature increase and a 40% augmentation in precipitation levels (see Jay *et al.* 2012, for a more extensive study). Admixture coefficients computed with the projected climatic variables are displayed in Fig-

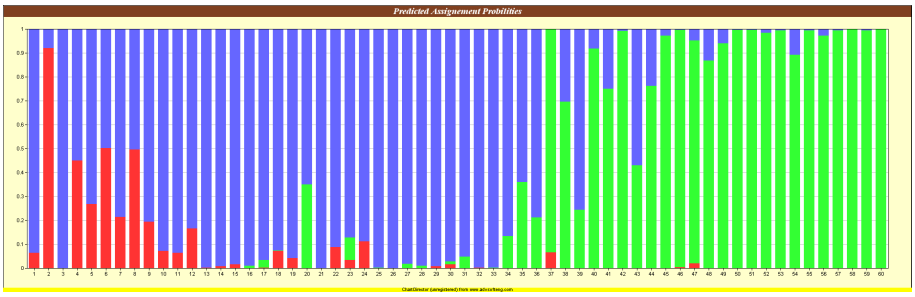


Figure 4: Admixture coefficients predicted for the 60 individuals contained in the test set. Admixture coefficients are computed for the 3 clusters inferred, using geographic, climatic and topographic information but no genetic data. The correlation between predicted admixture coefficients and coefficients estimated for the 60 test individuals using all environmental and genetic information is 0.85.

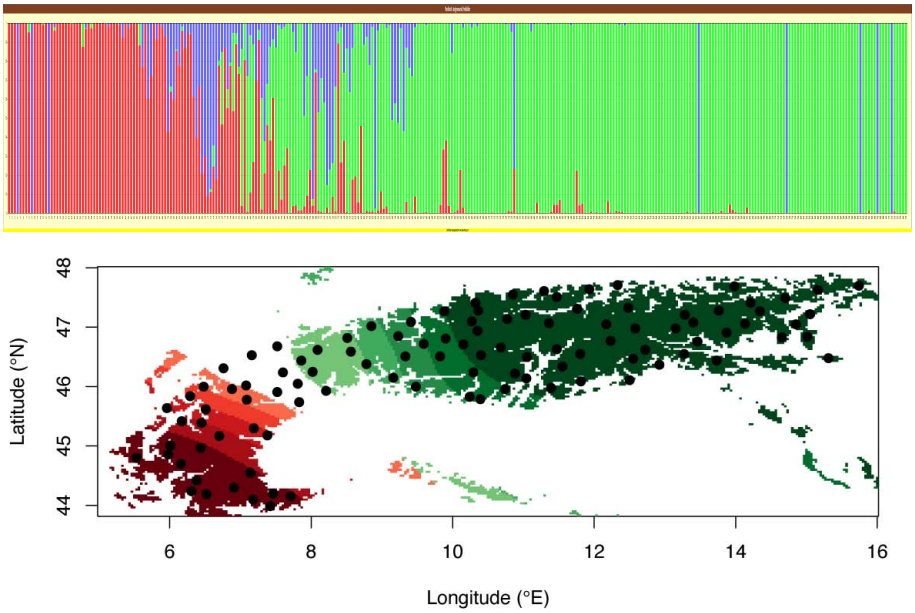


Figure 5: Admixture coefficients predicted for *Rhododendron ferrugineum L.* under a global warming scenario (2 °C temperature increase and 40% precipitation increase). A: Admixture coefficients are displayed on a bar chart. B: Admixture coefficients greater than 0.5 are displayed spatially using kriging methods.

ure 5. A comparison with current admixture coefficients (Figure 2) provides evidence that the contact zones between the central and the southern cluster, and between the central and the northern cluster, shift in the northward and westward directions respectively.

6. Conclusion

POPS is a software package that can be used to estimate population structure from individual multilocus genotypes and external covariates without assuming predefined populations.

POPS jointly infers population genetic structure and the effect of environmental covariates on this structure. **POPS** users can use environmental covariates to predict genetic structure or for changing environmental conditions. **POPS** can be used either from a graphical user interface or from a command-line engine. R scripts for post-processing results and displaying membership or admixture coefficients are available in the **POPS** package. **POPS** can be downloaded from http://membres-timc.imag.fr/Olivier.Francois/pops.html.

Acknowledgments

We are grateful to the anonymous reviewers for their critical reading and comments. We also thank Chibiao Chen for early work on the software implementation, and the INTRABIODIV consortium for providing the plant genetic data set. MGBB and OF acknowledge support from the persyvact project of the persyval-lab labex.

References

Aitken S, Yeaman S, Holliday J, Wang T, Curtis-McLane S (2008). "Adaptation, Migration or Extirpation: Climate Change Outcomes for Tree Populations." *Evolutionary Applications*, **1**(1), 95–111. doi:10.1111/j.1752-4571.2007.00013.x.

Albert J, Chib S (1993). "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association*, **88**(422), 669–679. doi:10.1080/01621459.1993.10476321.

Balding D (2006). "A Tutorial on Statistical Methods for Population Association Studies." *Nature Reviews Genetics*, **7**(10), 781–792. doi:10.1038/nrg1916.

Balkenhol N, Waits L, Dezzani R (2009). "Statistical Approaches in Landscape Genetics: An Evaluation of Methods for Linking Landscape and Genetic Data." *Ecography*, **32**(5), 818–830. doi:10.1111/j.1600-0587.2009.05807.x.

Bandeen-Roche K, Miglioretti D, Zeger S, Rathouz P (1997). "Latent Variable Regression for Multiple Discrete Outcomes." *Journal of the American Statistical Association*, **92**(440), 1375–1386. doi:10.1080/01621459.1997.10473658.

Besag J (1975). "Statistical Analysis of Non-Lattice Data." *The Statistician*, **24**(3), 179–195. doi:10.2307/2987782.

Chen C, Durand E, Forbes F, François O (2007). "Bayesian Clustering Algorithms Ascertaining Spatial Population Structure: A New Computer Program and a Comparison Study." *Molecular Ecology Notes*, **7**(5), 747–756. doi:10.1111/j.1471-8286.2007.01769.x.

Chung H, Flaherty B, Schafer J (2006). "Latent Class Logistic Regression: Application to Marijuana Use and Attitudes Among High-School Seniors." *Journal of the Royal Statistical Society A*, **169**(4), 723–743. doi:10.1111/j.1467-985x.2006.00419.x.

Davies N, Villablanca F, Roderick G (1999). "Determining the Source of Individuals: Multilocus Genotyping in Nonequilibrium Population Genetics." *Trends in Ecology and Evolution*, **14**(1), 17–21. doi:10.1016/s0169-5347(98)01530-4.

Dawson K, Belkhir K (2001). "A Bayesian Approach to the Identification of Panmictic Populations and the Assignment of Individuals." *Genetics Research*, **78**(01), 59–77. doi:10.1017/s001667230100502x.

Dayton C, Macready G (1988). "Concomitant-Variable Latent-Class Models." *Journal of the American Statistical Association*, **83**(401), 173–178. doi:10.2307/2288938.

Duminil J, Fineschi S, Hampe A, Jordano P, Salvini D, Vendramin G, Petit R (2007). "Can Population Genetic Structure Be Predicted from Life-History Traits?" *American Naturalist*, **169**(5), 662–672. doi:10.1086/513490.

Durand E, Jay F, Gaggiotti O, François O (2009). "Spatial Inference of Admixture Proportions and Secondary Contact Zones." *Molecular Biology and Evolution*, **26**(9), 1963–1973. doi:10.1093/molbev/msp106.

François O, Ancelet S, Guillot G (2006). "Bayesian Clustering Using Hidden Markov Random Fields in Spatial Population Genetics." *Genetics*, **174**(2), 805–816. doi:10.1534/genetics.106.059923.

Gugerli F, Englisch T, Niklfeld H, Tribsch A, Mirek Z, Ronikier M, Zimmermann N, Holderegger R, Taberlet P (2008). "Relationships Among Levels of Biodiversity and the Relevance of Intraspecific Diversity in Conservation – A Project Synopsis." *Perspectives in Plant Ecology, Evolution and Systematics*, **10**(4), 259–281. doi:10.1016/j.ppees.2008.07.001.

Intergovernmental Panel on Climate Change (2007). *Climate Change 2007: The Physical Science Basis*. Cambridge University Press.

Jakobsson M, Rosenberg N (2007). "**CLUMPP**: A Cluster Matching and Permutation Program for Dealing with Label Switching and Multimodality in Analysis of Population Structure." *Bioinformatics*, **23**(14), 1801–1806. doi:10.1093/bioinformatics/btm233.

Jay F, François O, Blum M (2011). "Predictions of Native American Population Structure Using Linguistic Covariates in a Hidden Regression Framework." *PLoS ONE*, **6**(1). doi:10.1371/journal.pone.0016227.

Jay F, Manel S, Alvarez N, Durand E, Thuiller W, Holderegger R, Taberlet P, François O (2012). "Forecasting Changes in Population Genetic Structure of Alpine Plants in Response to Global Warming." *Molecular Ecology*, **21**(10), 2354–2368. doi:10.1111/j.1365-294x.2012.05541.x.

Lee C, Mitchell-Olds T (2011). "Quantifying Effects of Environmental and Geographical Factors on Patterns of Genetic Differentiation." *Molecular Ecology*, **20**(22), 4631–4642. doi:10.1111/j.1365-294x.2011.05310.x.

Lichstein J, Simons T, Shriner S, Franzreb K (2002). "Spatial Autocorrelation and Autoregressive Models in Ecology." *Ecological Monographs*, **72**(3), 445–463. doi:10.2307/3100099.

Linzer D, Lewis J (2011). "**poLCA**: An R Package for Polytomous Variable Latent Class Analysis." *Journal of Statistical Software*, **42**(10), 1–29. doi:10.18637/jss.v042.i10.

Manel S, Schwartz M, Luikart G, Taberlet P (2003). "Landscape Genetics: Combining Landscape Ecology and Population Genetics." *Trends in Ecology and Evolution*, **18**(4), 189–197. doi:10.1016/s0169-5347(03)00008-9.

Nychka D, Furrer R, Sain S (2015). ***fields**: Tools for Spatial Data*. R package version 8.2-1, URL http://CRAN.R-project.org/package=fields.

Parmesan C, Yohe G (2003). "A Globally Coherent Fingerprint of Climate Change Impacts across Natural Systems." *Nature*, **421**(6918), 37–42. doi:10.1038/nature01286.

Pritchard J, Stephens M, Donnelly P (2000). "Inference of Population Structure Using Multilocus Genotype Data." *Genetics*, **155**(2), 945–959.

Qt Project (2015). *Qt: Cross-Platform Application Framework*. Version 5.5.0, URL http://www.qt.io/.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ripley B (1981). *Spatial Statistics*. John Wiley & Sons, New York. doi:10.1002/0471725218.

Segelbacher G, Cushman S, Epperson B, Fortin M, François O, Hardy O, Holderegger R, Taberlet P, Waits L, Manel S (2010). "Applications of Landscape Genetics in Conservation Biology: Concepts and Challenges." *Conservation Genetics*, **11**(2), 375–385. doi:10.1007/s10592-009-0044-5.

Sork V, Davis F, Westfall R, Flint A, Ikegami M, Wang H, Grivet D (2010). "Gene Movement And Genetic Association With Regional Climate Gradients In California Valley Oak (Quercus Lobata Née) In The Face Of Climate Change." *Molecular Ecology*, **19**(17), 3806–3823. doi:10.1111/j.1365-294x.2010.04726.x.

Spiegelhalter S, Best N, Carlin B, Linde A (2002). "Bayesian Measures of Model Complexity and Fit." *Journal of the Royal Statistical Society B*, **64**(4), 583–639. doi:10.1111/1467-9868.00353.

Stephens M (2000). "Dealing with Label Switching in Mixture Models." *Journal of the Royal Statistical Society B*, **62**(4), 795–809. doi:10.1111/1467-9868.00265.

Storfer A, Murphy M, Evans J, Goldberg C, Robinson S, Spear S, Dezzani R, Delmelle E, Vierling L, Waits L (2006). "Putting the Landscape in Landscape Genetics." *Heredity*, **98**(3), 128–142. doi:10.1038/sj.hdy.6800917.

Vermunt J, Magidson J (2003). "Latent Class Models for Classification." *Computational Statistics & Data Analysis*, **41**(3), 531–537. doi:10.1016/s0167-9473(02)00179-2.

Vounatsou P, Smith T, Gelfand A (2000). "Spatial Modelling of Multinomial Data with Latent Structure: An Application to Geographical Mapping of Human Gene and Haplotype Frequencies." *Biostatistics*, **1**(2), 177–189. doi:10.1093/biostatistics/1.2.177.

Affiliation:

Flora Jay
Muséum National d'Histoire Naturelle,
CNRS UMR 7206, Eco-Anthropologie et Ethnobiologie,
Université Paris Diderot
17 place du Trocadéro
75016 Paris, France
Telephone: +33/1/71214619
E-mail: flora.jay@mnhn.fr

Olivier François
Université Joseph Fourier, CNRS UMR 5525, TIMC-IMAG
Computational and Mathematical Biology
38042 Grenoble cedex, France
Telephone: +33/4/56520025
E-mail: olivier.francois@imag.fr
URL: http://membres-timc.imag.fr/Olivier.Francois/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
December 2015, Volume 68, Issue 9 *Submitted:* 2011-11-04
doi:10.18637/jss.v068.i09 *Accepted:* 2015-01-13